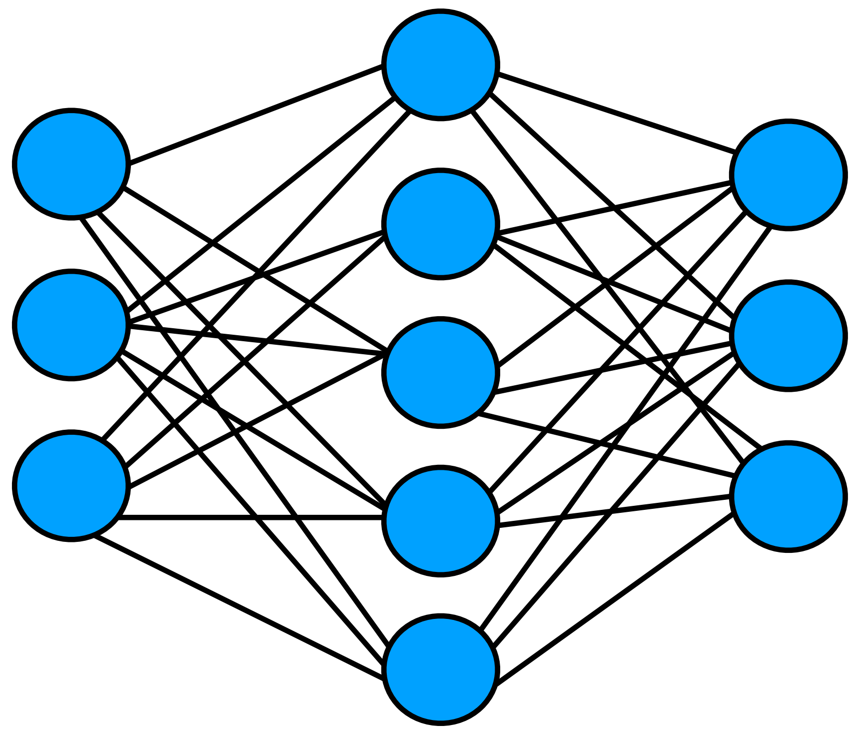
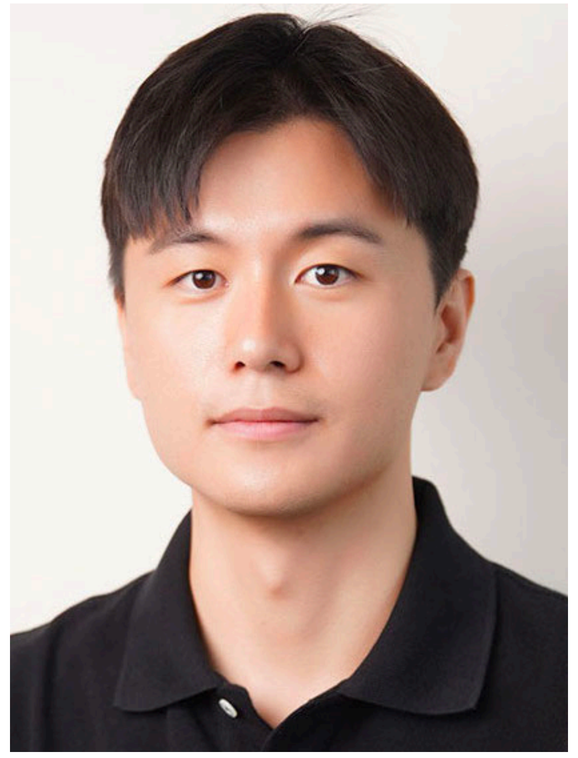
A Representer Theorem for Vector-Valued Neural Networks: Insights on Weight Decay Regularization and Widths of DNNs

Joe Shenouda
UW-Madison ECE



Rahul Parhi



Kangwook Lee

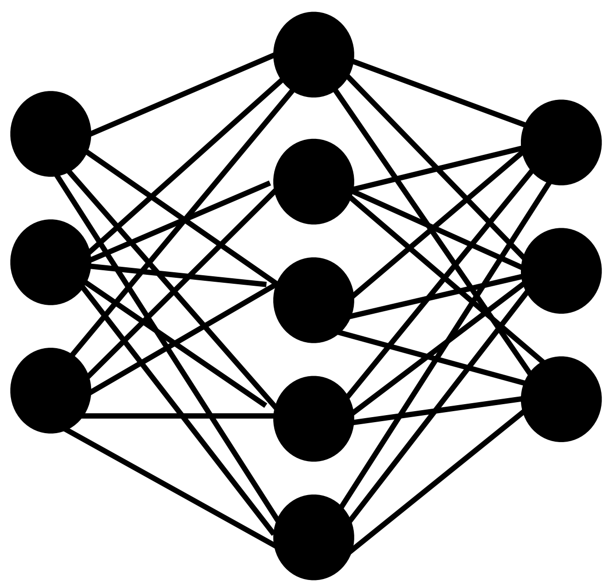
Rob Nowak

DP4ML ICML 2023

Motivations

- **Vector valued NNs** are key to understanding deep networks and multi-task learning.
- **Weight decay** is the most popular explicit regularizer for training deep neural networks (DNNs) and can have a **drastic effect** on the generalization ability of the network.

Background

$f_\theta =$  $\quad f_\theta : \mathbb{R}^d \to \mathbb{R}^D$

- Training a DNN with **weight decay** corresponds to minimizing a data fidelity loss plus the sum of squared weights.

Training with Weight Decay

$$\theta^{k+1} = \theta^k - \gamma \nabla_{\theta^k} \mathcal{L} - \boxed{\gamma \lambda \theta^k}$$

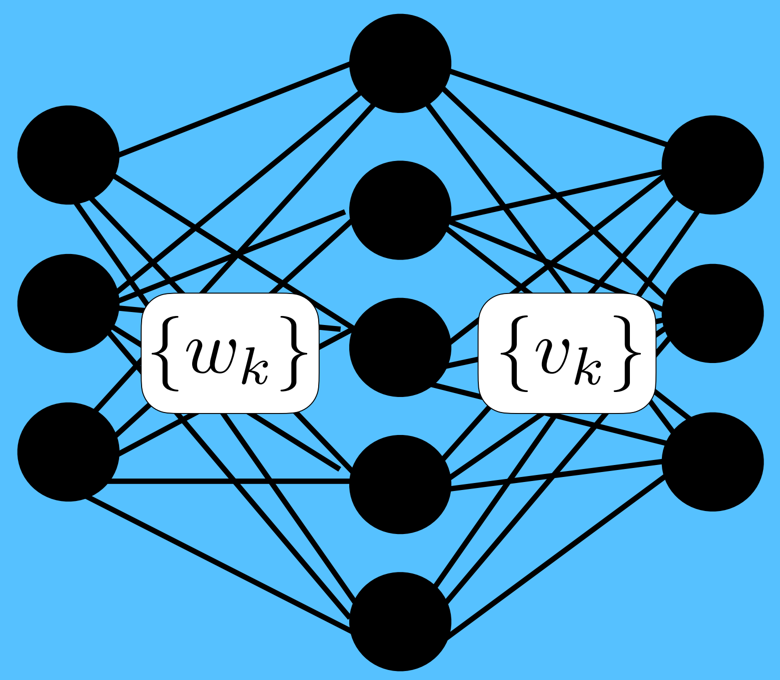
Weight Decay Objective

$$\min_\theta \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{y}_i, f(\boldsymbol{x}_i)) + \frac{\lambda}{2} \|\theta\|_2^2$$

Neural Balance Theorem

Theorem (NBT [Yang 2022, Parhi 2023])

Let f_θ be a function represented by a homogenous DNN that **solves the weight decay objective.**

$f_\theta =$ 

$$\min_{\{\boldsymbol{w}_k, \boldsymbol{v}_k\}} \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{y}_i, f(\boldsymbol{x}_i)) + \frac{\lambda}{2} \sum_{k=1}^{K} (\|\boldsymbol{w}_k\|_2^2 + \|\boldsymbol{v}_k\|_2^2)$$

Then for any **neuron** with input weight $\boldsymbol{w}_k$ and output weight $\boldsymbol{v}_k$ we have $\|\boldsymbol{w}_k\|_2 = \|\boldsymbol{v}_k\|_2$

Balanced!

The Secret Sparsity of Weight Decay

$$\min_{\{\boldsymbol{w}_k, \boldsymbol{v}_k\}_{k=1}^{\widehat{K}}} \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{y}_i, f(\boldsymbol{x}_i)) + \frac{\lambda}{2} \overbrace{\boxed{\sum_{k=1}^{\widehat{K}} \|\boldsymbol{w}_k\|_2^2 + \|\boldsymbol{v}_k\|_2^2}}^{\text{Weight decay}}$$

$$\min_{\{\boldsymbol{w}_k, \boldsymbol{v}_k\}_{k=1}^{\widehat{K}}} \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{y}_i, f(\boldsymbol{x}_i)) + \lambda \overbrace{\boxed{\sum_{k=1}^{\widehat{K}} \|\boldsymbol{v}_k\|_2 \|\boldsymbol{w}_k\|_2}}^{\text{Path-norm}}$$

$$\min_{\{\boldsymbol{w}_k, \boldsymbol{v}_k\}_{k=1}^{\widehat{K}}; \|\boldsymbol{w}_k\|_2 = 1} \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{y}_i, f(\boldsymbol{x}_i)) + \lambda \overbrace{\boxed{\sum_{k=1}^{\widehat{K}} \|\boldsymbol{v}_k\|_2}}^{\text{Group lasso}}$$

After Rebalancing

Secret Sparsity

Outline of Contributions

Vector-Valued Variation Spaces (VV Spaces)

- **Representer Theorem** showing shallow vector-valued NNs trained with weight decay solve data fitting problem over the VV Space.

Bounds on Necessary Width

- **Tighter bounds** on the necessary width for any homogenous layer of DNN, depending only on the **rank** of the internal features.

VV Spaces

Vector Valued Variation Spaces (VV Spaces)

Vector Measure

$$VV = \left\{ f(\boldsymbol{x}) = \int_{\mathbb{S}^d} \sigma(\boldsymbol{w}^T \boldsymbol{x}) \boxed{d\nu(\boldsymbol{w})} : \|f\|_{VV} < \infty \right\}$$

$$\|f\|_{VV} = \sum_k \|v_k\|_2$$ **(For finite-width networks)**

Representer Theorem

Also extends to deep neural networks

For any dataset $\{x_i, y_i\}_{i=1}^N$ and any lower semicontinuous loss function $\mathcal{L}$ there exists a solution to,

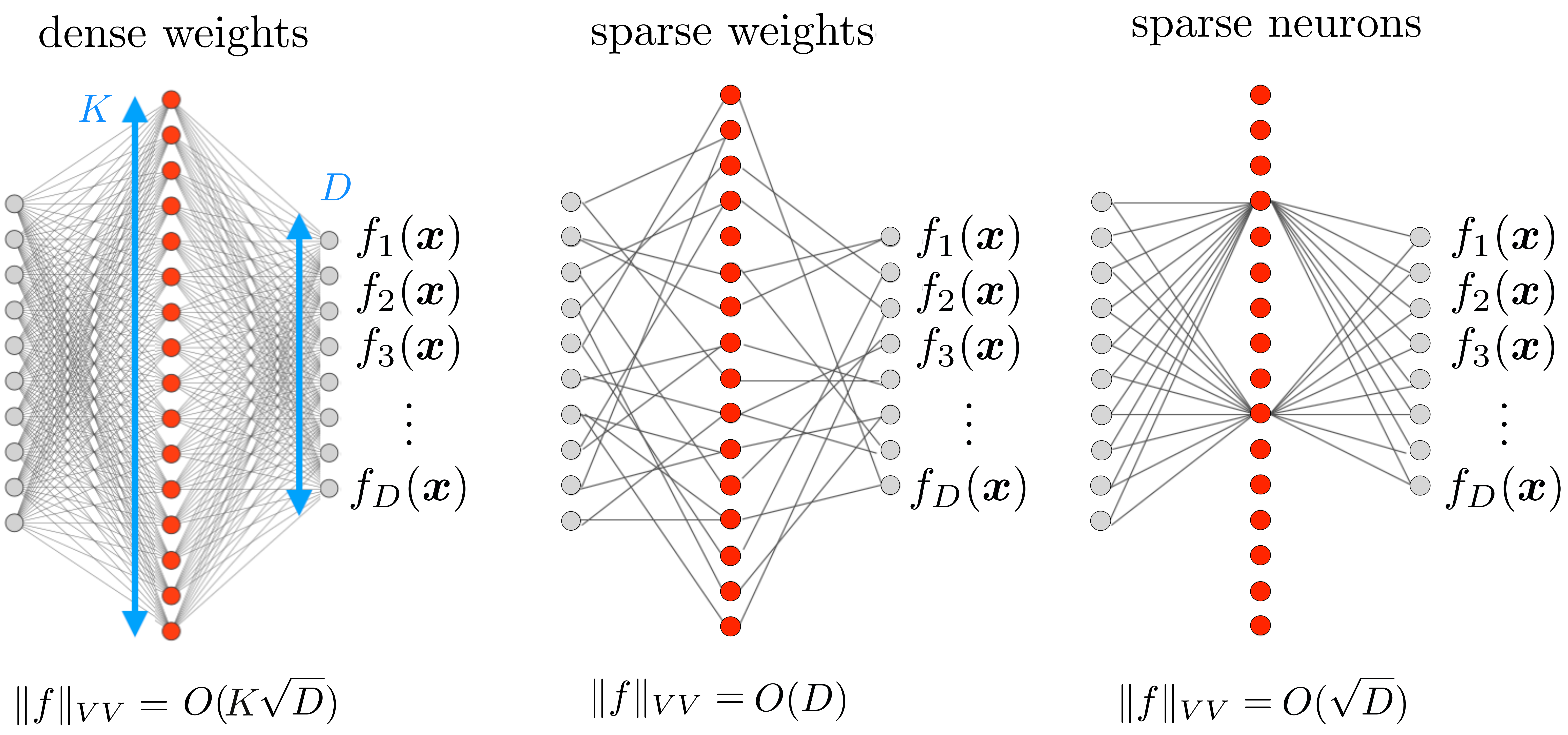
$$\arg\min_{f \in VV} \sum_{i=1}^{N} \mathcal{L}(y_i, f(x_i)) + \lambda \|f\|_{VV}$$

with the following **representation**

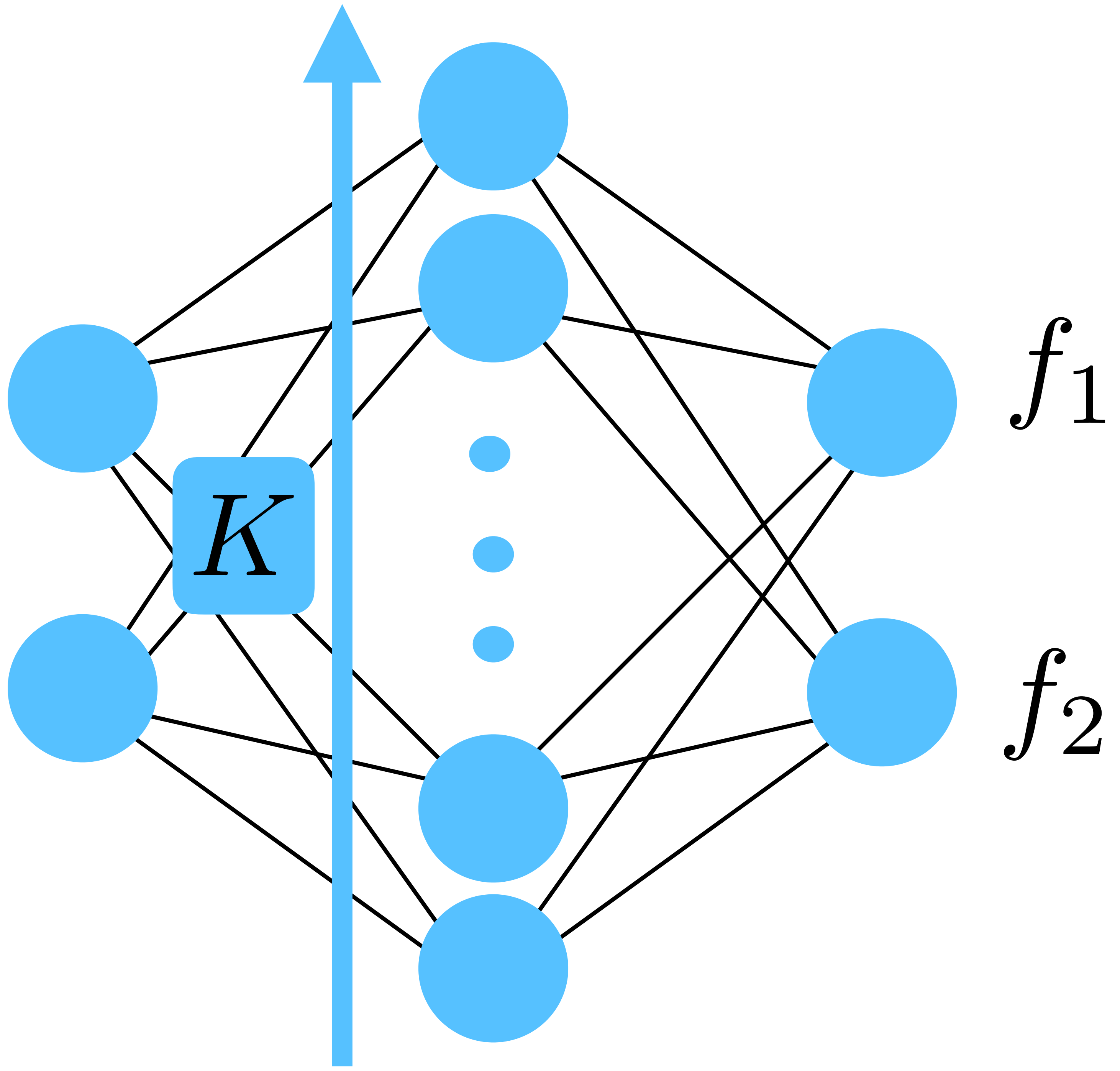
$$f_{\theta^*}(x) = \sum_{k=1}^{K_0} \boldsymbol{v}_k \sigma\left(\boldsymbol{w}_k^T \boldsymbol{x}\right)$$

$$\boxed{K_0 \leq N^2}$$

The dual space of vector-valued continuous functions is the space of vector-valued measures.

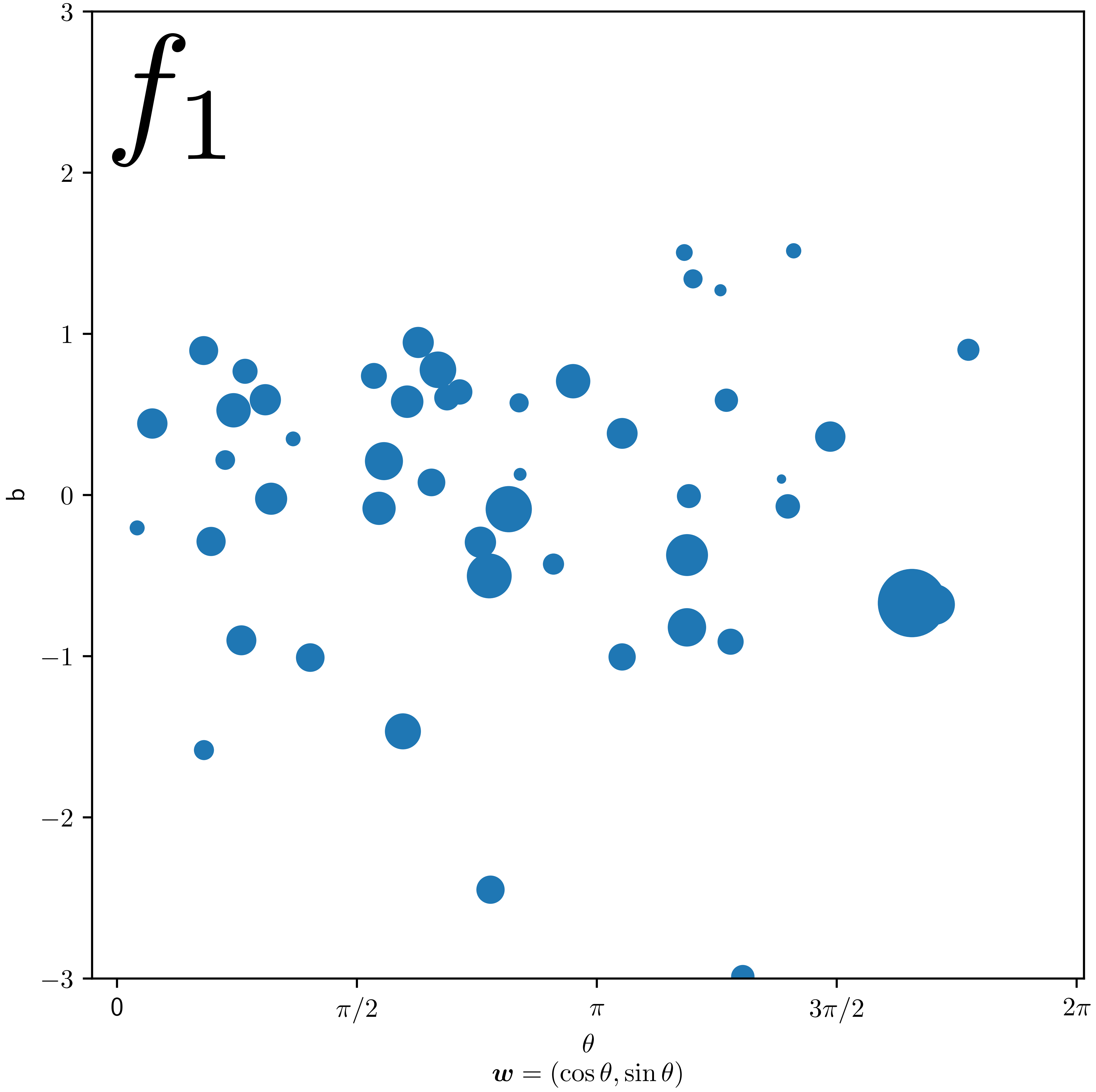

dense weights
sparse weights
sparse neurons
K
D
$f_1(\boldsymbol{x})$
$f_2(\boldsymbol{x})$
$f_3(\boldsymbol{x})$
$\vdots$
$f_D(\boldsymbol{x})$
$\|f\|_{VV} = O(K\sqrt{D})$
$\|f\|_{VV} = O(D)$
$\|f\|_{VV} = O(\sqrt{D})$

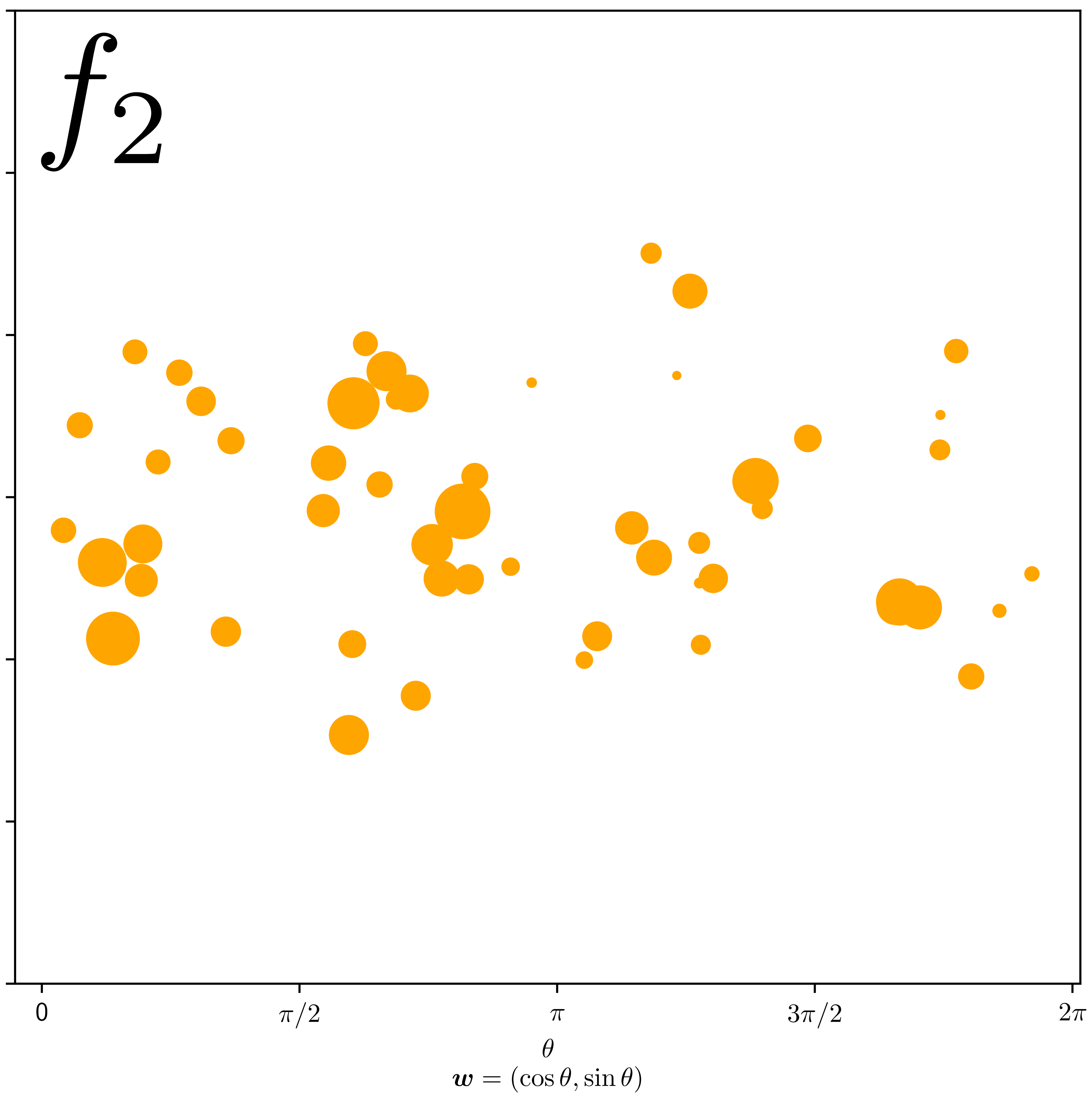
Simple Experiment



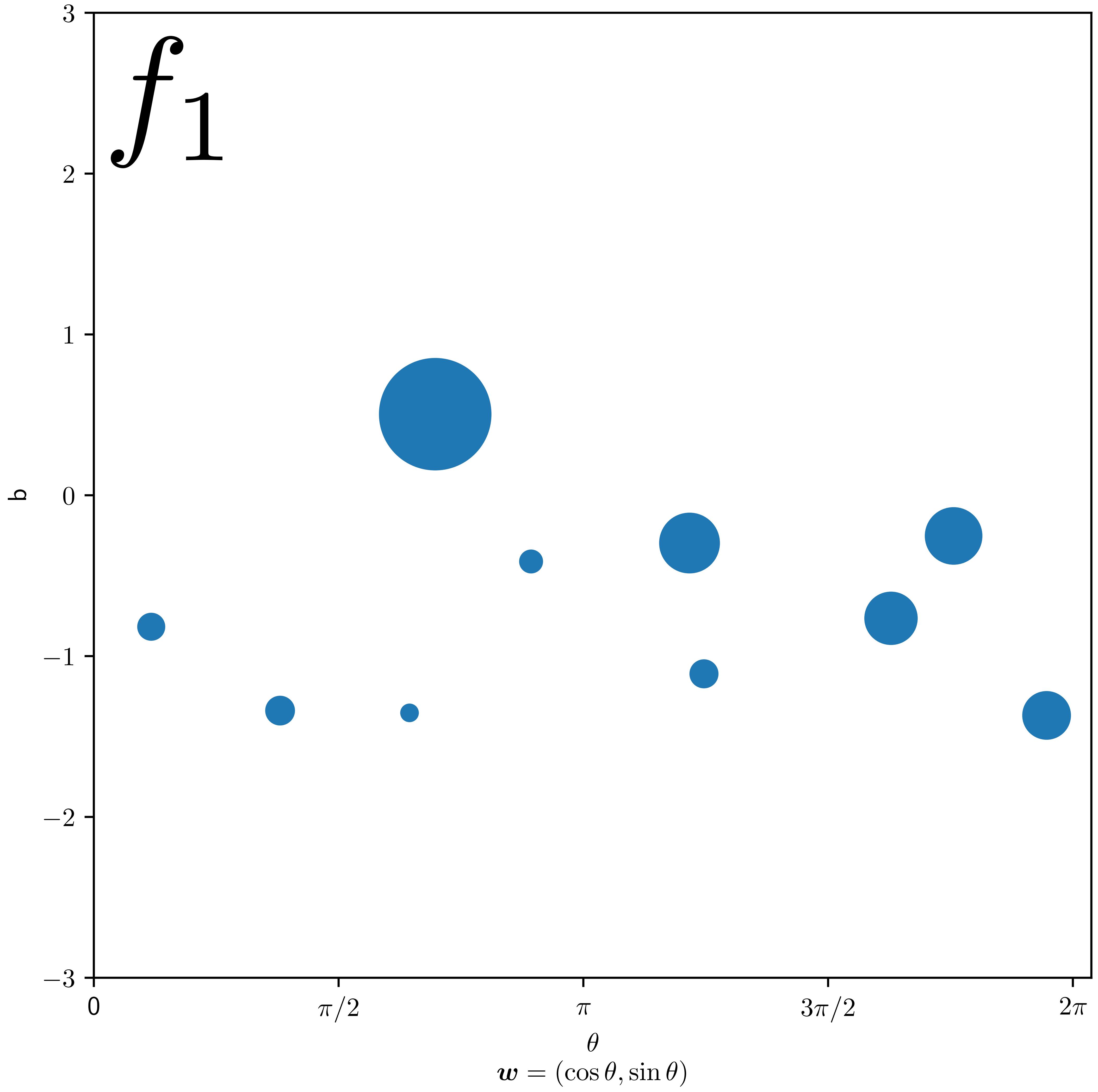
$$f(\boldsymbol{x}) = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \sum^{K} \boldsymbol{v}_k \sigma(\boldsymbol{w}_k^T \boldsymbol{x} - b_k)$$

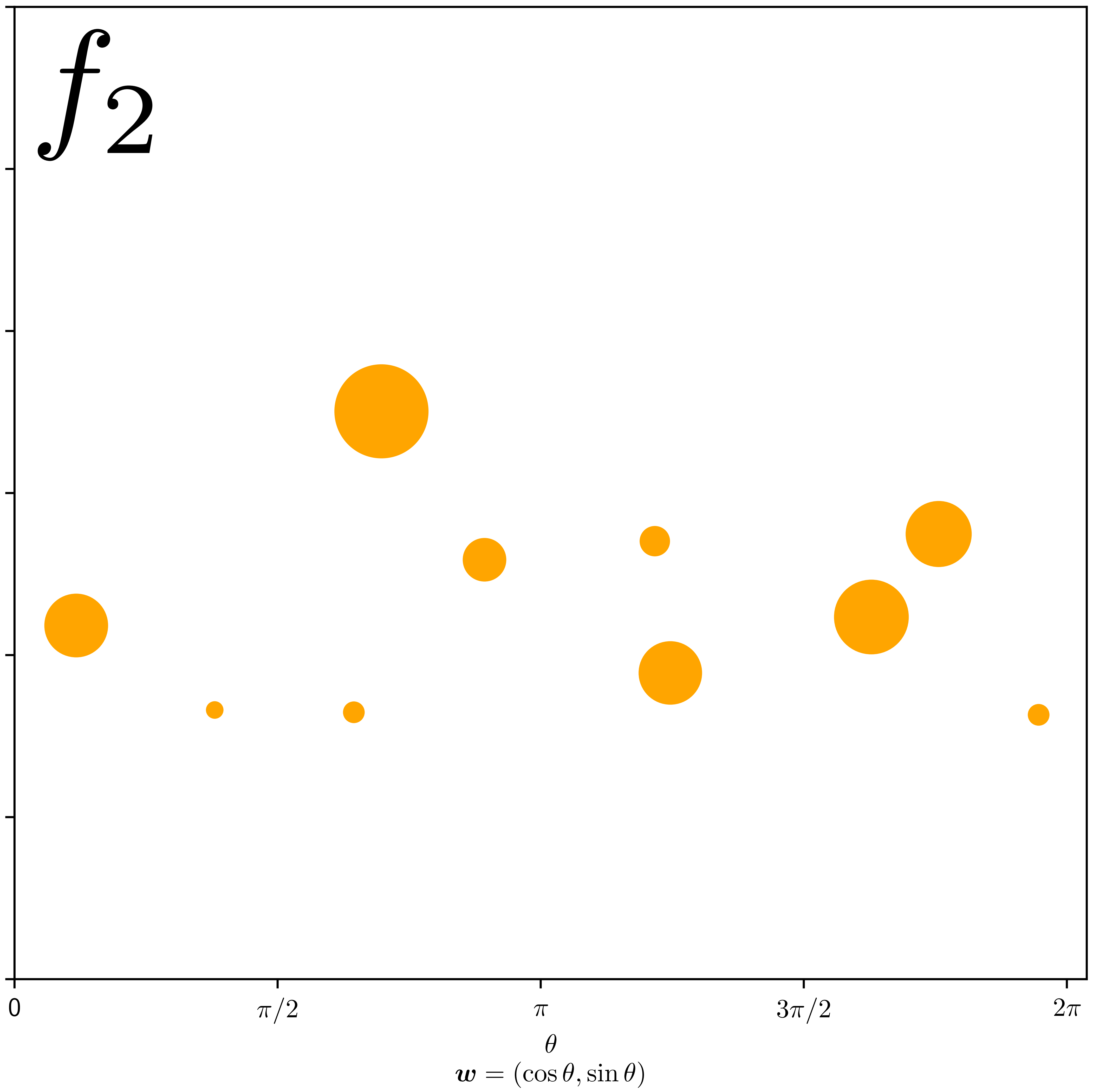
No Weight Decay —> No Sharing





Weight Decay —> Neuron Sharing

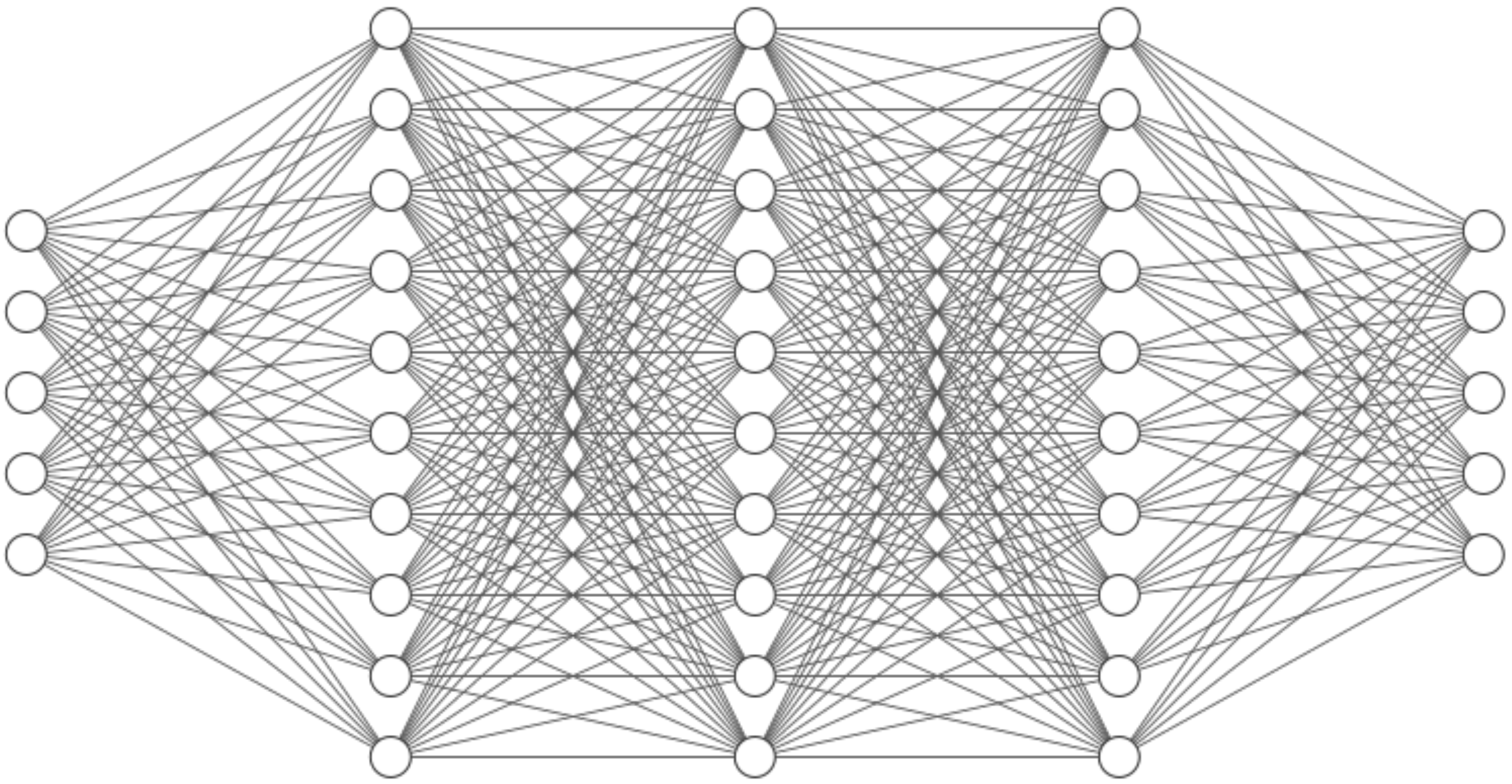




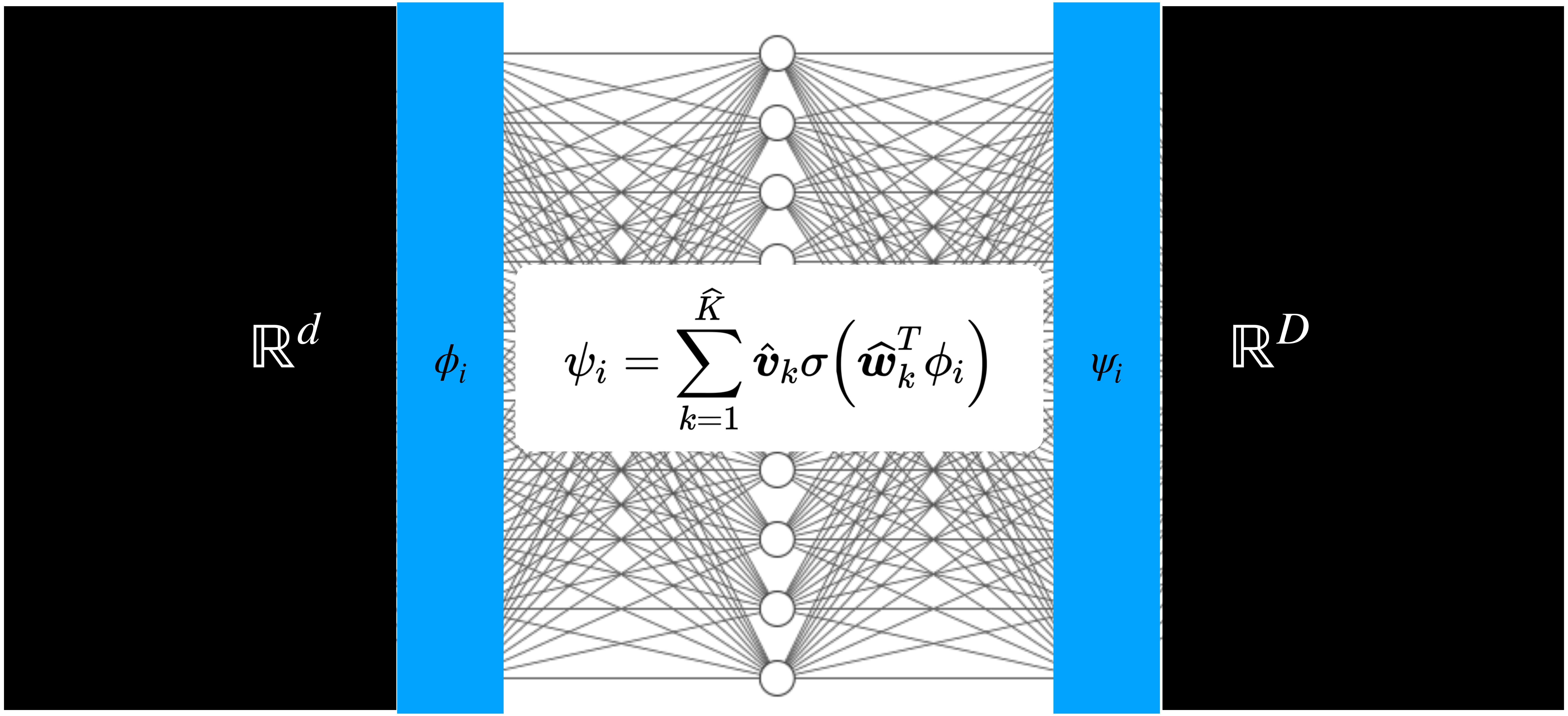
Tighter Bounds on Necessary Widths of DNNs

Bound on Network Width

- Suppose f_θ is a **DNN** which solves the weight decay objective



Bound on Network Width



$$\widetilde{\Phi} = \left[\sigma\left(\widehat{\boldsymbol{w}}_k^T \phi_1\right) \cdots \sigma\left(\widehat{\boldsymbol{w}}_k^T \phi_N\right)\right]$$

$$\boldsymbol{\Psi} = [\psi_1 \quad \psi_2 \cdots \psi_N]$$

Main Result

Assume f_θ is a DNN which solves the **weight decay objective**,

for any **homogenous layer** with input features ϕ_i and output features ψ_i, such that

$$\psi_i = \sum_{k=1}^{\widehat{K}} \hat{\boldsymbol{v}}_k \sigma\left(\widehat{\boldsymbol{w}}_k^T \phi_i\right) \qquad i = 1, \cdots, n$$

There exists another **optimal** representation of the form

$$\psi_i = \sum_{k=1}^{K} \boldsymbol{v}_k \sigma\left(\boldsymbol{w}_k^T \phi_i\right) \qquad i = 1, \cdots, n$$

Where,

$$\boxed{K \leq \operatorname{rank}(\Psi) \cdot \operatorname{rank}(\widetilde{\Phi})}$$

Questions?